Lecture 7: Mean-field neural networks

Summary of the class so far: (Mid-semester "sanity check")

2 central questions: { Why optimization is tractable?
Why do NNs generalize?

Two hypotheses:

* **Trackability via overparametrization:** the highly non-convex landscape simplifies dramatically as # parameters $\gg$ # samples so that local search methods (e.g., GD/SGD) succeed at finding global minima ("interpolating" solution)

* **Implicit regularization of training algorithm:** algo itself has the selection of a solution that generalizes well on new data DESPITE the model being overparametrized and interpolating noisy data.

Lectures 2/3/4: Overparametrization is not incompatible with generalization

- **Lecture 2:** capacity of NNs (Rademacher complexity) depends on norms / spectral properties of the weights and not # parameters.
- **Lecture 3/4:** solution selected by GD (min-norm interpolating solut°) in linear models can overfit benignly thanks to a self-induced regularization

This motivates the study of gradient-trained NNs that are highly-overparametrized, and as a first approximation # neurons $\to \infty$

2-layer NN: $f(x;\theta) = \frac{1}{\sqrt{M}} \sum_{j=1}^{M} a_j \sigma(\langle w_j, x \rangle)$ $\|x\|_2 = 1$

$a_j^0 \overset{iid}{\sim} N(0,1)$ $w_j^0 \overset{iid}{\sim} N(0, Id_d)$

"Standard Xavier initialization" ("used" in pytorch)

(This is more subtle than this → see next week)

Lecture 5: As $M \to \infty$, GD on this NN effectively behave as a linear model

$$f(x;\theta^t) \approx f(x;\theta^0) + \langle \theta^t - \theta^0, \nabla_\theta f(x;\theta^0) \rangle \quad (3)$$

CV exponentially fast to global solution (provably trained efficiently)

↳ "Lazy regime" $\|\theta^t - \theta^0\|_2 \ll 1$

(general mechanism that has nothing to do with NNs a priori)

Lecture 6: Linearized NNs are kernel methods which have limited adaptivity (adapt to "smoothness" but not to other low-dimensional structure)

→ "fixed feature" "fixed representation" method $\nabla_\theta f(x;\theta_0)$

→ suffer from the curse of dimensionality

If θ^t moves away from θ^0 during dynamics, can hope to learn good representation $\nabla_\theta f(x,\theta^t)$ and do much better

"feature learning", "representation learning", "learning a good kernel"

Example: Learning a single neuron

Consider $y_i = \sigma(\langle w_*, x_i \rangle) + \varepsilon_i$ $\quad x_i \sim \mathrm{Unif}(S^{d-1})$, $\|w_*\|_2 = \sqrt{d}$

Fit with $f(x;\Theta) = \frac{\alpha}{\sqrt{M}} \sum_{j=1}^{M} a_j \sigma(\langle w_j, x \rangle)$

GD on $\hat{R}_n(\Theta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i;\Theta))^2$

Thm [Montanari, Zhong, '22, Ghorbani, Mei, Misiakiewicz, Montanari '21]

Fix $\delta > 0$ and take $Md \geq n^{1+\delta}$. Let $\hat{\Theta}$ be the solution found in the Lazy training regime ($\alpha \gtrsim \sqrt{n^2/Md}$).

If $n \leq d^{\ell+1-\delta}$, then

$$\liminf_{n,d \to \infty} R(f(\cdot;\hat{\Theta})) \geq \|P_{>\ell}\sigma\|_{L^2}^2$$

$P_{>\ell}\sigma$: fit at most a d^ℓ polynomial approximat° to σ

This is disappointing given that $\sigma(\langle w_*, x \rangle)$ is

extremely simple and we can fit with one neuron !!!
The problem is w_j° don't move during optimization in the lazy regime and we are trying to fit $\sigma(\langle w_*, x\rangle)$ with $\sigma(\langle w_j^\circ, x\rangle)$

↳ in high dimensions $\sup_{j\in[M]} \frac{|\langle w_j^\circ, w_*\rangle|}{\|w_j^\circ\|_2 \|w_*\|_2} \lesssim \frac{1}{\sqrt{d}}$

If instead $M = 1$ $\quad f(x;\Theta) = \sigma(\langle x, w\rangle)$

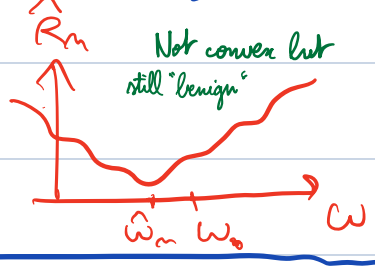
$$\hat{R}_n(w) = \frac{1}{n}\sum_{i=1}^{n} (y_i - \sigma(\langle x_i, w\rangle))^2$$

Thm: [Bai, Mei, Montanari, '18] Assume σ bounded and 3 times differentiable bounded derivatives, and $\sigma'(t)$ for all $t\in\mathbb{R}$. Then there exists a constant $C>0$ such that if $n \geq C\, d \log d$, we have

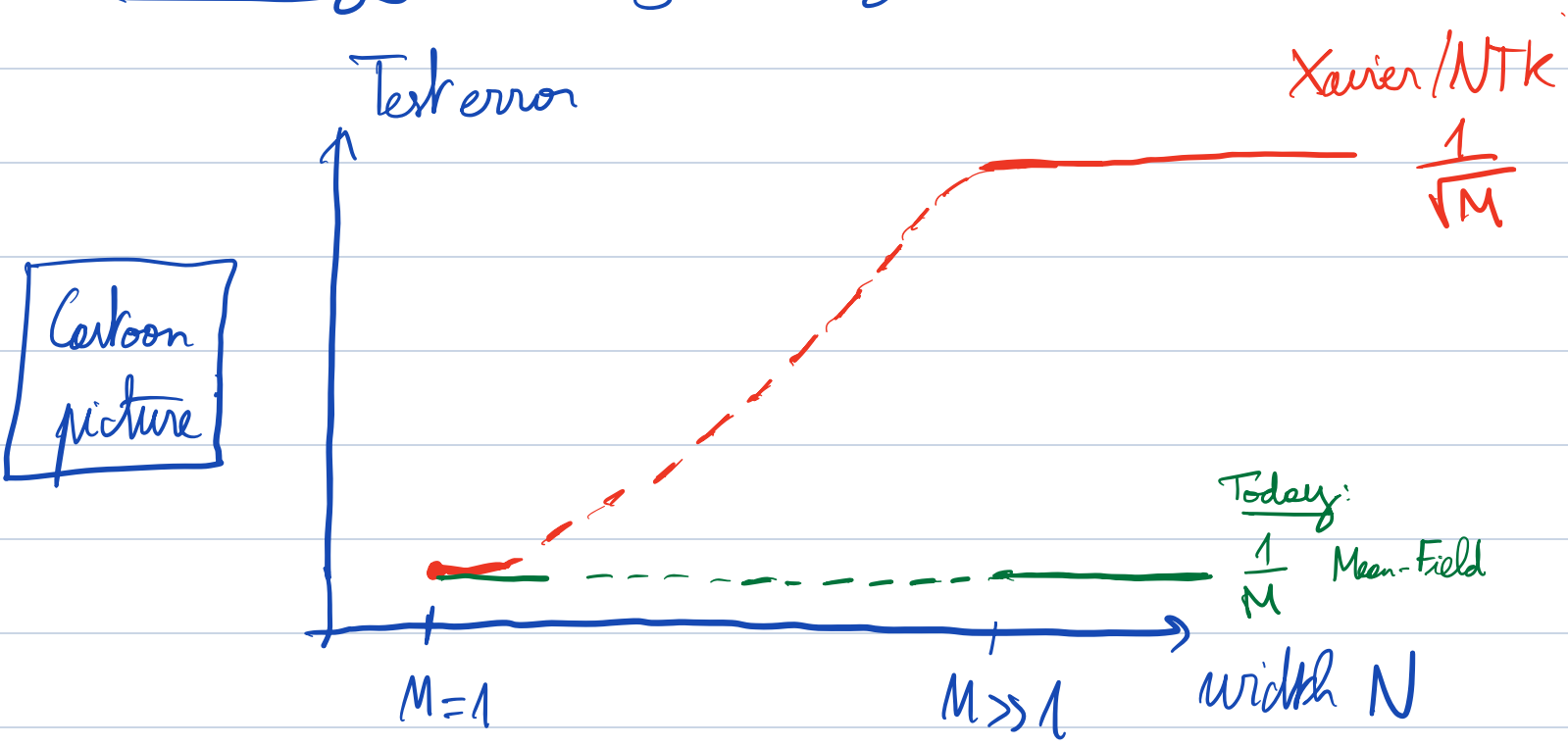
(i) $\hat{R}_n(w)$ has unique local/global minimizer $\hat{w}_n$

(ii) GD cv to $\hat{w}_n$

(iii) $R(\hat{w}_n) \leq C\sqrt{\frac{d\log n}{n}}$



Summary: learning a single neuron



Cartoon picture

Test error deteriorates as $M \to \infty$. What is going wrong?

→ This is only a specific infinite-width limit

↳ can chose different scalings and obtain other limits with different behavior.

Today: Mean-Field limit; scaling $\frac{1}{M}$

Mean-Field scaling

$$f(x;\Theta) = \frac{1}{M}\sum_{j=1}^{M} a_j\,\sigma(\langle w_j, x\rangle)$$

$\frac{\alpha_M}{\sqrt{M}}$ $\qquad$ $\alpha_M := \frac{1}{\sqrt{M}}$

Why is this a natural parametrization?

1. Lecture 1: approximation theory of NNs

$$\frac{1}{M}\sum_{j=1}^{M} a_j\,\sigma(\langle w_j, x\rangle) \xrightarrow[M\to\infty]{} \int a(w)\,\sigma(\langle w, x\rangle)\,\mu(dw)$$

f_* well approximated with $a(w)\mu(dw)$ $\longrightarrow$ then sample

2. Lecture 2: Rademacher complexity $\lesssim r_0\sqrt{\frac{d}{n}}$ $\qquad \frac{\|a\|_1}{M} \le r_0$ with MF scaling.

[With NTK scaling $\lesssim r_0\sqrt{\frac{Md}{n}}$ $\qquad \frac{\|a\|_1}{M} \le r_0$]

3. Lecture 5: $\frac{\alpha_M}{\sqrt{M}}$ lazy regime $\alpha_M \gtrsim \sqrt{\frac{n^2}{Md}}$

To not be in the lazy regime as $M\to\infty$ for n/d fixed

we need $\alpha_M \lesssim \frac{1}{\sqrt{M}}$ as $M \to \infty$

Rmk: Systematic study of infinite width limits of NNs through "ABC"-parametrization [Greg Yang, 2021]

→ classify scalings that give stable and non-trivial limits as $M \to \infty$

→ MF is special among these limits (for 2-layer) as it maximizes "feature learning"

(Might cover it next week... very useful (easy) computation)

Stochastic Gradient Descent (SGD)

Let's rewrite $f(x;\Theta) = \frac{1}{M}\sum_{j=1}^{M} \sigma(x;\theta_j)$ $\quad \theta_j \in \mathbb{R}^D$

$$\Theta = (\theta_1, \dots, \theta_M) \in \mathbb{R}^{MD}$$

e.g. : $\sigma(x;\theta_j) = a_j \tilde{\sigma}(\langle x, w_j \rangle)$ $\quad \theta_j = (a_j, w_j) \in \mathbb{R}^{d+1}$

Consider squared loss : $\ell(y, \hat{y}) = \frac{1}{2}(y-\hat{y})^2$

Test error : $R_D(\Theta) = \frac{1}{2} \mathbb{E}_{(y,x)\sim D}\left[(y - f(x;\Theta))^2\right]$

Train error : $\hat{R}_n(\Theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}(y_i - f(x_i;\Theta))^2$

$= R_{\hat{D}_n}(\Theta)$ $\quad \hat{D}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{y_i, x_i}$

Simplest algorithm: Gradient Descent (≈ 1850)

$$\Theta^{k+1} = \Theta^k - \eta \nabla \hat{R}_n(\Theta^k)$$

≈ 1950: we do not need to compute exact gradient

$$w^{k+1} = w^k - \eta\left(\nabla \hat{R}_n(w^k) + z^k\right)$$

z^k: iid noise with zero mean

→ noise averages out and algo converges as GD (with suitable step sizes)

SGD (with batch size = 1)

* Initialize: w^0 $\quad \theta_j^0 \sim_{iid} \ell_0$

* At each step: sample (y_k, x_k)

$$w^{k+1} = w^k - \eta M \nabla_w \ell(y_k, f(x_k, w^k))$$

scaling that gives $\Theta_M(1)$ updates

Two cases (1) Finite training samples (y_n, x_n) drawn uniformly at random from $\hat{D}_n$ "multi-pass SGD"

as $\eta \to 0$ w^t follows GF on $\hat{R}_n(w)$

$$\frac{d}{dt} w^t = -M \nabla_w \hat{R}_n(w^t)$$

(2) Infinite training samples $(y_k, x_k) \overset{iid}{\sim} \mathcal{D}$

"one-pass SGD" → each sample only used once

as $\eta \downarrow 0$ $\frac{d}{dt}\Theta^t = -M \nabla_\Theta R(\Theta^t)$ "population GF"

Both cases can be treated in a unified way $\hat{R}_n = R_{\hat{\mathcal{D}}_n}$ (population dist is $\hat{\mathcal{D}}_n$)

↳ below "one pass SGD"

Dynamics • $\theta_j^0 \overset{iid}{\sim} \rho_0$

$$\bullet\ \theta_j^{k+1} = \theta_j^k + \eta \nabla_{\theta_j} \sigma(x_k; \theta_j)\left(y_k - \frac{1}{M}\sum_{s=1}^{M} \sigma(x_k; \theta_s)\right)$$

Goal: we want to produce a simplified description of the SGD dynamics in this scaling

* $\eta \downarrow 0$: $\frac{d}{dt}\Theta^t = -M \nabla_\Theta R(\Theta^t)$

* $M \to \infty$: $f(x; \Theta^t) = f(x; \hat{\rho}_t^{(M)})$

$$f(x; \rho) = \int \sigma(x; \theta)\, \rho(\theta) \qquad \hat{\rho}_t^{(M)} = \frac{1}{M}\sum_{j=1}^{M} \delta_{\theta_j^t}$$

Empirical weight distribution

Note that $\theta_j^0 \overset{iid}{\sim} \rho_0$ at $t=0$ $\quad \hat{\rho}_0^{(M)} \Rightarrow \rho_0$

"weak convergence of measures"

It is reasonable to expect

$$\hat{\rho}_t^{(M)} \Rightarrow \rho_t \quad \text{for all } t \geq 0.$$

[i.e., $\forall$ bounded C^0 fct $\int f \, d\hat{\rho}_t \to \int f \, d\rho_t$]

where ρ_t is a deterministic measure following some PDE which we will describe next.

Rmk: This is classically referred to as "propagation of chaos": as $M \to \infty$, a system of interacting particles can be well approximated by independent particles interacting with a "mean-field" ρ_t

Mean-field limit : static properties

Risk: $R_M(\Theta^{(M)}) = \frac{1}{2}\mathbb{E}_{xy}\left[\left(y - f(x;\Theta^{(M)})\right)^2\right]$

$$= \frac{1}{2}\left[R_{\#} + \frac{2}{M}\sum_{j=1}^{M} V(\theta_j) + \frac{1}{M^2}\sum_{i,j=1}^{M} U(\theta_i,\theta_j)\right]$$

where

$$R_{\#} = \mathbb{E}[y^2] \qquad V(\theta) = -\mathbb{E}_{yx}\left[y\,\sigma_*(x;\theta)\right]$$

$$U(\theta_1,\theta_2) = \mathbb{E}_x\left[\sigma_*(x;\theta_1)\,\sigma_*(x;\theta_2)\right]$$

Interpretation: $R_M(\Theta^{(M)})$ is the energy of a system of M particles $\theta_j \in \mathbb{R}^D$, interacting with an external potential $V(\theta)$ and via a pairwise potential $U(\theta_1,\theta_2)$

In particular: U is P.S.D. ie for any bounded and compactly supported function h, we have

$$\int h(\theta_1)\,U(\theta_1,\theta_2)\,h(\theta_2)\,d\theta_1\,d\theta_2 \geq 0$$

"repulsive interaction".

Gradient w.r.t one particle:

$$M \nabla_{\theta_j} R_M(\Theta) = \nabla_{\theta_j} \left(V(\theta_j) + \frac{1}{M} \sum_{i=1}^{M} U(\theta_j, \theta_i) \right)$$

$$=: \nabla_{\theta_j} \Psi(\theta_j ; \Theta)$$

As M is large, it is natural to replace $\theta_1, \dots, \theta_M$ by a density $\rho \in \mathcal{P}(\mathbb{R}^D)$

$$R(\rho) = \frac{1}{2} R_{\#} + \int V(\theta) \rho(d\theta) + \int U(\theta_1, \theta_2) \rho(d\theta_1) \rho(d\theta_2)$$

And $\Psi(\theta; \rho) = V(\theta) + \int U(\theta, \theta') \rho(d\theta')$

$$" = \frac{\delta R(\rho)}{\delta \rho(\theta)} "$$

("additional energy of adding a single particle at $\theta \in \mathbb{R}^D$")

Rmk: $R(\hat{\rho}^{(M)}) = R_M(\Theta)$

$$\frac{d}{dt} \theta_j^t = -\nabla_\theta \Psi(\theta_j ; \hat{\rho}_t^{(M)})$$

(this is a completely equivalent descript° of GF on M neurons)

Hence, heuristically, $\hat{\rho}_t^{(M)} \Rightarrow \rho_t$

"consistency equation"

Each particle $\theta^t \sim \rho_t$ satisfy $\frac{d}{dt} \theta^t = -\nabla_\theta \Psi(\theta^t ; \rho_t)$

PDE characterization

We can describe evolution of ρ_t in terms of a PDE

Below is a heuristic derivation: assume (has a density) $\rho_t(d\theta) = \rho_t(\theta)\, d\theta$ (for simplicity)

Consider a differentiable test function $g: \mathbb{R}^D \to \mathbb{R}$

$$\frac{d}{dt}\mathbb{E}[g(\theta^t)] = \frac{d}{dt}\int g(\theta)\,\rho_t(\theta)\,d\theta$$

(differentiating under integral)

$$= \int g(\theta)\,\partial_t \rho_t(\theta)\,d\theta$$

On the other hand, exchanging derivative / expectation

$$\frac{d}{dt}\mathbb{E}[g(\theta^t)] = \mathbb{E}\left[\frac{d}{dt}g(\theta^t)\right] = \mathbb{E}\left[\left\langle \nabla_\theta g(\theta^t), \frac{d}{dt}\theta^t \right\rangle\right]$$

$$= -\mathbb{E}\left[\left\langle \nabla_\theta g(\theta^t), \nabla_\theta \psi(\theta^t_t; \rho_t)\right\rangle\right]$$

$$= -\int \left\langle \nabla_\theta g(\theta), \nabla_\theta \psi(\theta; \rho_t)\right\rangle \rho_t(\theta)\,d\theta$$

Integrate by part

$$= \int g(\theta)\, \underbrace{\nabla\cdot}_{\Delta}\left[\rho_t(\theta)\nabla_\theta \psi(\theta;\rho_t)\right] d\theta$$

(divergence $\nabla_\theta \cdot v(\theta) = \sum_{i=1}^{D} \frac{\partial}{\partial \theta_i} v_i(\theta)$

Hence putting equalities together:

$$\int g(\theta) \left\{ \partial_t \rho_t(\theta) - \nabla \cdot [\rho_t(\theta) \nabla_\theta \psi(\theta; \rho_t)] \right\} d\theta = 0$$

"Weak form" of the following PDE

$$\boxed{\partial_t \rho_t = \nabla \cdot [\rho_t \nabla_\theta \psi(\theta; \rho_t)]}$$

This is often referred to as "Distributional Dynamics"

This is known as a McKean-Vlasov type PDE.

PDE independently proven in 4 concurrent work [CB'18, MMN'18, RVE'18, SS'18]

If initialize a particle $\theta_j^0 \sim \rho_0$ and

$\frac{d}{dt} \theta_j^t = -\nabla_\theta \psi(\theta_j^t; \rho_t)$ then $\theta_j^t \sim \rho_t$ for all $t \geq 0$

Approximation guarantees

How well $f(x;\rho_t)$ tracks $f(x;\Theta^k)$?

where • Θ^k solut° of SGD with M neurons
η step size

• ρ_t solut° of above PDE.

Assumptions : (i) $|\sigma|, |y| \le C$, $\nabla_\theta \sigma(x;\theta)$ C-subGaussian

(ii) $V(\theta)$, $U(\theta_1,\theta_2)$ have bounded Lipschitz gradient

(weaker assumpt° in the paper)

Thm [Mei, Misiakiewicz, Montanari, '19]

There exists a constant K that only depends on assumptions s.t. for all $T \ge 0$, with probability $\ge 1 - e^{-z}$

$$\sup_{k \in [0, \frac{T}{\eta}] \cap \mathbb{N}} \left| R_M(\Theta^k) - R(\rho_{k\eta}) \right| \le \frac{K e^{KT}}{\sqrt{M}} \left(\sqrt{\log M} + z \right)$$

error due to $M \to \infty$ approx

$$+ K e^{KT} \left(\sqrt{D + \log M} + z \right) \sqrt{\eta}$$

error due to $\eta \to 0$ approx.

Rmk:

- $\hat{\ell}_k^{(M)} \approx \ell_k \eta \qquad t = k\eta$
- For T fix, $M \gtrsim 1$ and $\eta \lesssim \frac{1}{D}$ for MF to be a good approximation.
  - → In particular M does not need to depend on D but only on properties of the target fct
  - → similar to approximat° theory (e.g Barron's result)
- e^{kT} is bad but in some sense unavoidable
  - ↳ $T = O(1) \qquad \frac{T}{\eta} = O(D)$ SGD steps
  - ↳ good approx limited to this linear scaling
  - ↳ will talk more about it in following lectures

Wasserstein gradient flow description

The PDE has an interesting structure: it is a gradient flow in the space of distributions.

It is worth fleshing out this relation as it connects MF limit to a rich field of mathematics: optimal transport.

What is a gradient flow?

Consider:

* (M, ρ) a metric space (ρ = distance)
* $F: M \to \mathbb{R}$

Trajectory: $z(0), z(\varepsilon), z(2\varepsilon), \dots z(k\varepsilon), \dots$

$$z(k\varepsilon + \varepsilon) = \underset{z \in M}{\operatorname{argmin}} \left\{ F(z) + \frac{1}{2\varepsilon} \rho(z, z(k\varepsilon)) \right\}$$

As $\varepsilon \searrow 0$, this defines a continuous trajectory $z(t)$

"GF on F in (M, ρ)"

We minimize "movements" in (M, ρ) to descend on $F(z)$

E.g. Euclidean space $\frac{dz(t)}{dt} = -\nabla F(z(t))$ is indeed the limit of the above

Previous PDE is a Gradient flow on $R(\rho)$ in the metric space $(\mathcal{P}_2(\mathbb{R}^D), W_2)$

* $\mathcal{P}_2(\mathbb{R}^D)$: space of probs ρ on $\mathbb{R}^D$ with $\int \|\theta\|_2^2 \rho(d\theta) < \infty$

* W_2: Wasserstein distance

$$W_2(\mu,\nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \left(\int \|x-y\|_2^2 \, \gamma(dx,dy) \right)^{\frac{1}{2}}$$

where $\Gamma(\mu,\nu)$ is the set of all couplings of μ and ν

i.e. joint probs on $\mathbb{R}^{D\times D}$ with marginals μ and ν

$$\rho_{t+\varepsilon} \approx \operatorname*{argmin}_{\rho \in \mathcal{P}(\mathbb{R}^D)} \left\{ R(\rho) + \frac{1}{2\varepsilon} W_2(\rho,\rho_t)^2 \right\}$$

In particular $\dfrac{dR(\rho_t)}{dt} = -\int \|\nabla \varphi(\theta;\rho_t)\|_2^2 \, \rho_t(d\theta)$.

≤ 0

Global convergence

$R(\rho_t)$ is monotonically decreasing. Can we show global convergence? (And have some success as lazy regime)

$$\lim_{t \to \infty} R(\rho_t) = \inf_{\rho} R(\rho)$$

[Here we consider one-pass SGD hence global CV is directly on test error, i.e., directly implies that we generalize well]

→ not quite: showing global CV of this Wasserstein flow is hard and there are many subtle difficulties.

Below I describe some basic properties.

Static properties:

Minimizing $R_M(\textcircled{u})$ is not much different that minimizing $R(\rho)$

Lemma: Assume $\exists \varepsilon, K > 0$ such that

$$R(\varrho) \leq \inf_{\varrho} R(\varrho) + \varepsilon \Rightarrow \int U(\theta, \theta)\, \varrho(d\theta) \leq K$$

Then:

$$\left| \inf_{\underline{\theta} \in \mathbb{R}^{Md}} R_M(\underline{\theta}) - \inf_{\varrho} R(\varrho) \right| \leq \frac{K}{M}$$

Proof: Take ϱ_* s.t. $R(\varrho_*) \leq \inf R(\varrho) + \delta \qquad \delta \leq \varepsilon$

(in particular $\int U(\theta,\theta)\, \varrho_*(d\theta) \leq K$ by assumption)

Consider $(\theta_j)_{j \in [M]} \overset{\text{iid}}{\sim} \varrho_*$, then simple calculation yields

$$\mathbb{E}_{\varrho_*}\left[R_M(\underline{\theta})\right] - R(\varrho_*) = \frac{1}{M}\left\{\int U(\theta,\theta)\varrho_*(d\theta) - \underbrace{\int U(\theta_1,\theta_2)\,\varrho_*(d\theta_1)\varrho_*(d\theta_2)}_{\geq 0 \text{ by PSD}}\right\}$$

$$\leq \frac{1}{M}\int U(\theta,\theta)\,\varrho_*(d\theta) \leq \frac{K}{M}$$

Hence $\inf_{\underline{\theta}} R_M(\underline{\theta}) \leq \inf_{\varrho} R(\varrho) + \frac{K}{M} + \delta$ for any δ □

We further have the following characterization of minimizer

Prop: Assume U, V are C^0 and bounded from below.
Assume $\exists \varepsilon, \gamma, K > 0$, s.t.

$$R(\rho) \le \inf_{\rho} R(\rho) + \varepsilon \implies \int \|\theta\|^\gamma \rho(d\theta) \le K$$

Then (1) $\inf_{\rho} R(\rho)$ is achieved at some ρ_*.

(2) ρ_* is a minimizer iff

$$\mathrm{supp}(\rho_*) \subseteq \mathrm{argmin}_\theta \Psi(\theta; \rho_*)$$

(Proof is by doing a perturbation of ρ_*)

Corollary: Suppose ρ_* satisfy (i) $\mathrm{supp}(\rho_*) = \mathbb{R}^D$ and (ii) $\Psi(\theta; \rho_*)$ is a constant
Then ρ_* is a minimizer.

Proof: This is direct from previous proposition (2)

Dynamical properties:

ρ is a fixed point of the PDE if and only if

$$(*) \quad \text{supp}(\rho) \subseteq \{\theta : \nabla_\theta \Psi(\theta; \rho) = 0\}$$

(simply comes from $\frac{d}{dt} R(\rho_t) = -\int \|\nabla_\theta \Psi\|_2^2 \rho_t(d\theta)$)

There will be ∞ number of "bad" stationary points

(e.g. any stationary pt of landscape with finite # of neurons will be a stationary of the PDE with empirical measure $\hat{\rho}^{(M)}$)

It is hard & subtle to argue that ρ_t will not get trapped in these stationary points.

Here is a simple result:

Thm: Assume (a) $\rho_t \to \rho_\infty \in \mathcal{P}_2(\mathbb{R}^D)$

(b) $\text{supp}(\rho_\infty) = \mathbb{R}^D$

(+ U, V differentiable with bounded gradient)

Then ρ_∞ is a minimizer, that is

$$R(\rho_t) \to \min_\rho R(\rho)$$

Proof: ρ_∞ stationary point, i.e

$$\mathbb{R}^D = \mathrm{supp}(\rho_\infty) \subseteq \{\theta : \nabla\Psi(\theta; \rho_\infty) = 0\}$$

hence $\Psi(\theta; \rho_\infty)$ cste and we conclude using the corollary above □

Assumption $\mathrm{supp}(\rho_\infty) = \mathbb{R}^D$ is very restrictive

[Chizat, Bach, '18] [Wojtowytsch, '22]
[Nguyen, Pham, '20]

Always need to assume $\begin{cases} \rho_t \to \rho_\infty & (*) \\ \rho_\infty \text{ has some abstract properties} \end{cases}$

[(*) this is not a benign condition as ρ_t might oscillate while $R(\rho_t) \searrow R_*$]

Nonetheless: these proofs are instructive and show that it is hard to keep a distribut° in a non-optimal stationary point
(tractability via overparametrizat°)

Summary:
- No quantitative general CV guarantees
- Even qualitative guarantees are under abstract assumptions

Noisy SGD

Consider the following variant of SGD

→ add $\frac{\lambda}{2}\|\Theta\|_2^2$ regularization to the risk

→ add at each step Gaussian noise to the gradient

At each step: $\Theta_j^{k+1} = (1-\lambda\eta)\Theta_j^k + \eta \nabla_\theta \sigma(x;\Theta_j^k)(y_k - f(x_k;\Theta^k)) + \sqrt{\frac{\eta}{\beta}}\, g_j^k$

$\rightarrow g_j^k \overset{iid}{\sim} N(0, I_D)$

Risk: $R_\lambda(\rho) = R(\rho) + \frac{\lambda}{2}\int \|\theta\|_2^2 \rho(d\theta)$

$$\Psi_\lambda(\theta;\rho) = \Psi(\theta;\rho) + \frac{\lambda}{2}\|\theta\|_2^2$$

Then as $M \to \infty$ $\eta \downarrow 0$

The resulting dynamics

$$\partial_t \rho_t = \nabla_\theta \cdot [\rho_t \nabla_\theta \Psi_\lambda(\theta;\rho_t)] + \frac{1}{\beta}\Delta \rho_t$$

Additional diffusion term

"inverse temperature"

Laplacian

$$\Delta \cdot f(\theta) = \sum_{i=1}^{D} \frac{\partial^2}{\partial \theta_i^2} f(\theta)$$

This is also a Wasserstein gradient flow but now on the free energy

$$F_{\lambda,\beta}(\rho) = R_\lambda(\rho) - \frac{1}{\beta} S(\rho)$$

with $S(\rho) = -\int \rho(\theta) \log \rho(\theta) \, d\theta$ the entropy of ρ

The interesting property when adding entropic regularizer is that the Free energy is strongly convex and there is only one stationary point which is the global minimizer of $F_{\lambda,\beta}(\rho)$

Lemma: For any $\lambda \geq 0$ and $\beta > 0$, $F_{\lambda,\beta}(\rho)$ has at most one fixed point $F_{\lambda,\beta}(\rho_*) < \infty$.
Furthermore ρ_* has density satisfying self-consistent Boltzmann equation

$$\rho_*(\theta) = \frac{1}{Z(\lambda,\beta)} \exp\left(-\beta \mathcal{V}_\lambda(\theta; \rho_*)\right)$$

Proof: This follows from showing that the dissipat° equation

$$\frac{d F_{\lambda,\beta}(\rho_t)}{dt} = -\int_{\mathbb{R}^D} \left\| \nabla_\theta \left[\Psi_\lambda(\theta;\rho_t) + \frac{1}{\beta} \log \rho_t(\theta) \right] \right\|_2^2 \rho_t(\theta)\, d\theta$$

It is $= 0$ iff ρ_* satisfy equation above

In particular this means that

$$\lim_{t\to\infty} F_{\lambda,\beta}(\rho_t) = \min_{\rho} F_{\lambda,\beta}(\rho)$$

If β taken sufficiently small $\min_{\rho} F_{\lambda,\beta}(\rho) \approx \min_{\rho} R_\lambda(\rho)$

[We need $\beta = \Omega(D)$]

We get: $\lim_{t\to\infty} R(\rho_t) \leq \inf_{\rho} R(\rho) + \tilde{O}\left(\frac{D}{\beta}\right)$

Great!!! Global CV as long as $\beta < \infty$

but what about convergence time?

Standard approach: prove Log-Sobolev inequality for this setting

→ If LSI with LS constant α then

$$F_{\lambda,\beta}(\rho_t) - F_{\lambda,\beta}(\rho_*) \leq e^{-C\frac{\alpha}{\beta}t}\left(F_{\lambda,\beta}(\rho_0) - F_{\lambda,\beta}(\rho_*)\right)$$

However: best we can show $\alpha \geq \lambda\beta\, e^{-C\beta}$

Hence if $\beta = \Theta(\mathbb{D})$ $\quad \alpha \gtrsim e^{-\Theta(\mathbb{D})}$

and require $T = e^{\Theta(\mathbb{D})}$ to converge

→ this cannot be improved in general (i.e., there exist settings where we will need $e^{\mathbb{D}}$ time to CV).

Returning to single neuron learning example

We return to the example of learning a single neuron which we started with

This illustrate how the PDE can be simplified in settings with symmetries.

assume $\sigma'(t) > 0 \quad \forall t \in \mathbb{R}$

Consider: $x_i \sim N(0, I_d) \qquad y_i = \sigma(\langle w_*, x_i \rangle) + \varepsilon_i$

Model: $f(x; \theta) = \int a\, \sigma(\langle w, x \rangle)\, \rho(da, dw)$

Training: "uninformative initialization"

$$\rho_0(da, dw) = P_A \otimes N\left(0, \frac{\gamma^2}{d} I_d\right)$$

P_A: initialization of a $\perp$ initialization of w

$$\partial_t \rho_t = \nabla \cdot \left[\rho_t \nabla_\theta \Psi(\theta; \rho_t) \right]$$

Exploiting symmetries: data distribution (y, x) remain invariant under a rotation that leaves ω_* unchanged

Hence: if $R \in \mathbb{R}^{d\times d}$ s.t $R\omega_* = \omega_*$

then $R_\# \rho_t$ is a solution of the PDE with initialization $R_\# \rho_0$

Here $R_\# \rho_0 = \rho_0 \rightarrow$ by uniqueness of the solution

we must have $R_\# \rho_t = \rho_t$ for all $t \geq 0$.

Hence ρ_t is invariant under rotations that leave $R\omega_* = \omega_*$

i.e only depend on $(a, \underbrace{\langle \omega_*, \omega \rangle}_{s}, \underbrace{\|P_\perp \omega\|_2}_{r})$

$$\rho_t(da\, d\omega) = \bar{\rho}_t(da, ds, dr) \otimes \mathrm{Unif}(\mathbb{S}^{d-2})$$

Denote $z = (a, s, r)$

$$\Psi(z; \bar{\rho}_t) = \Psi(\theta; \rho_t) = V(\theta) + \int U(\theta, \theta') \rho_t(d\theta')$$

$$V(\theta) = -\mathbb{E}[y\, a\, \sigma(\langle \omega, x \rangle)] \qquad U(\theta, \theta') = \mathbb{E}[a a' \sigma(\langle \omega, x \rangle) \sigma(\langle \omega', x \rangle)]$$

$$V(\theta) = -a\, \mathbb{E}\left[\sigma(\langle \omega_0, x\rangle)\, \sigma(\langle \omega, x\rangle)\right]$$

$$= -a \underset{\substack{G_1, G_2 \\ \sim N(0,1)}}{\mathbb{E}} \left[\sigma(G_1)\, \sigma(s G_1 + r G_2)\right]$$

$$V(\theta,\theta') = aa' \underset{\substack{G_1,G_2,G_3 \\ \sim N(0,1)}}{\mathbb{E}}\left[\sigma(s G_1 + r G_2)\, \sigma\left(s' G_1 + \frac{\langle \omega_\perp, \omega'_\perp\rangle}{r} G_2 + \sqrt{r'^2 - \frac{\langle \omega_\perp, \omega'_\perp\rangle^2}{r^2}}\, G_3\right)\right]$$

When taking integral over $\omega'_\perp = r' u \quad u \sim \mathrm{Unif}(\mathbb{S}^{d-2})$
this indeed only depend on (a, s, r).

Hence the dynamics only depends on 3 parameters instead of $d+1$
Reduced form $\bar{\rho}_t \in \mathcal{P}(\mathbb{R}^3)$

$$\partial_t \bar{\rho}_t = \partial_a\left(\bar{\rho}_t\, \partial_a \Psi(z; \bar{\rho}_t)\right) + \partial_s\left(\bar{\rho}_t\, \partial_s \Psi(z; \bar{\rho}_t)\right) + \frac{1}{r}\partial_r\left(r \bar{\rho}_t\, \partial_r \Psi(z; \bar{\rho}_t)\right)$$

Initialization translates into $\bar{\rho}_0 = \underbrace{P_{A}}_{a^0} \otimes \underbrace{N\left(0, \frac{\gamma^2}{d}\right)}_{s^0} \otimes \underbrace{Q_{d-1,\gamma}}_{r^0}$

where $Q_{k,\gamma} = \frac{\gamma}{\sqrt{k+1}} \sqrt{\chi^2_{(k)}}$

The above PDE can be solved numerically efficiently

$$\lim_{t \to \infty} \underbrace{R(\varrho_t)}_{\text{excess test error}} = 0$$

For $\varepsilon > 0$, $\exists\ T = T(\varepsilon)$ indep of d s.t. $R(\varrho_t) \le \varepsilon$

Then taking $M \gtrsim \frac{e^{kT(\varepsilon)}}{\varepsilon^2} = \Theta_d(1)$

$$\eta \lesssim \frac{e^{-kT(\varepsilon)}}{\varepsilon^2 d} = \Theta_d\left(\frac{1}{d}\right)$$

We get $R_M(\Theta^{\frac{I}{\eta}}) \le R(\varrho_T) + \varepsilon \le 2\varepsilon$.

Total # samples $= \Theta_d(d)$ # neurons $= \Theta_d(1)$

(Success?)

Conclusion

Mean-Field scaling: $\rho_t \neq \rho_0$ "feature learning"

From single neuron example: can indeed adapt to low-dim structure and vastly outperform NTK.

Goal of $M \to \infty$ is to simplify analysis of SGD.
Did we succeed?

Pros:
- Compact description of dynamics in terms of PDE
- When setting has symmetries, can exploit them to reduce the effective dimension of the parameters

Cons:
- These PDE are hard to analyze in general
- Hard to prove qualitative/quantitative CV guarantees
- only good approx for $\mathcal{O}(1)$ steps of SGD

Ongoing line of work!!!